Applied Data Science Capstone Project: Full Report

Author: Sudeep Gupta

1. Business Problem:

The Tourism industry of New York is one of the most prominent industries in the city, quite understandably so as it is a cultural melting pot and gets millions of visitors from all parts of the world. Several local tourism agencies have sprung up in the different areas of the city and they compete against each other for providing the most memorable experience of New York to the enthusiastic tourists. In order to bring about such an experience, these agencies have to keep abreast of the changing dynamics of the popular spots in the city such as staying up to date with what restaurants are becoming the most popular in a locality or what museums are pulling in the most visitors. This helps them in suggesting the best and most fulfilling places to their visitors. Furthermore, they also have to figure out which localities are similar in nature and cluster different tourist places together so that their packages can include localities exhibiting varied culture and tastes in order to give a complete as well as cosmopolitan experience of New York to their customers.

To solve all these challenges and much more, these tourism agencies can leverage the powers of Data Science, especially the services of the Foursquare API which can answer the most important queries of their customers instantly. Moreover, the insights cultivated from the analysis of the tourism data of these places shall help such agencies to re-invent their business strategies and truly establish their hegemony over fellow competitors.

2. Data Requirements:

For the above problem, we would mainly focus on leveraging the regular as well as a few premium call services of the Foursquare API to extract the data required to analyse and study the tourist traffic and other significant details about the popular spots in New York. The data would consist of facts such as how many people frequent a given tourist spot in a day and which place receives the maximum number of visitors. It would also specify the type of tourists a particular place receives which can be leveraged to match tourists of a certain ethnicity/taste to the places that they would most likely want to visit during their stay in New York.

For instance, if a certain Mexican restaurant is famous in some locality of New York, tourists who would be interested in trying Mexican food can be matched up to that restaurant. In the same way, we can use the data obtained from the Foursquare API to match other places like museums, clubs etc to potential visitors.

Furthermore, the data extracted shall also contain the exact location of the tourist spot in New York i.e. its latitude and longitude. Similar spots which are close to each other can be clubbed together and suggested to the tourists saving both their time as well as money. This information can also be utilized to prepare tourist bus routes and specialized packages for the visitors, optimizing their tour and providing an overall seamless experience.

3. Methodology:

To perform an exploratory analysis of all the neighbourhoods in New York City and the tourism potential that they offer, the Foursquare API request was utilized to obtain the most frequented venues for each of the neighbourhoods. First of all, a data frame was created containing information about the total number of venues, geographical coordinates of venues and other details grouped in terms of each neighbourhood was created. This data frame gave an initial insight into which neighbourhoods had the most tourism potential.

This data frame was then utilized to create a bar graph consisting of the top 30 neighbourhoods in New York City which had the largest number of most popular neighbourhoods. Exploratory data analysis was done in each of these neighbourhoods and the exact number of most frequented venues along with other venue details was generated. This information can be vital for any tourism agency as it gives them a direct view into which places and destinations in the city promise the most revenue.

Secondly, we went into the process of iterating through all the neighbourhoods obtained in the data frame and performed an investigative analysis into the top 5 venues for each neighbourhood. This was a pretty interesting process as it showed what places attracted the most tourists for any given locality in New York. From Ice-Cream parlours to massage centres to Afghan Restaurants, we got to see some very interesting places thrown up in this query. Furthermore, the exact frequency of the tourists coming into these destinations was recorded which gave a deep insight into their popularity and ability to attract visitors.

Thirdly, we also performed an exploratory analysis into the rankings of the most popular venues grouped according to their respective neighbourhoods. This data might come handy for the agencies while planning their priorities for choosing destinations at a given locality in New York City. The rankings would provide a great parametric to include or eliminate potential places from the itineraries of their patrons. Furthermore, it gives an interesting read on which common places are popular in most of the neighbourhoods.

Machine Learning approach to cluster localities of similar Nature:

We utilized the K-nearest neighbour (KNN) Machine Learning algorithm to cluster the neighbourhoods into groups of similar kinds. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression). We utilized a classification based KNN on our New York neighbourhood's dataset to group similar neighbourhoods together.

This classification provided us with extremely valuable insights into how we can cluster tourist areas of a similar nature together to offer an exotic, culturally rich tourism package to the visitors. Surely, the various competing tour & travel agencies of New York would look at the results of the KNN with huge interest as it might help them to re-innovate their business strategies and assist them in offering tourists customised packages.

4. Results:

This data science project definitely produced some pretty captivating results when it comes to the behavioural patterns of the New York tourism flow and popularity of the kinds of places that tourists like to frequent the most. We got to see how some lesser known localities in the Big Apple such as **Murray Hill, Chelsea, Yorkville, Greenpoint** etc. showcased exciting tourism potential by topping the charts of the places with the most popular venues in New York City.

Secondly, we analysed the types of the most common venues in each of the neighbourhoods in the city of New York and obtained pretty satisfactory results. We got to know that the olden goldies of Tourism centres such as **Coffee shops, food joints, ice-cream parlours, Pizzerias** etc still dominate the tourism game in pretty much all the parts of the Big Apple. Thus, giving these places their due importance and investing in them would always pay off for the tourism agencies.

Thirdly and most importantly, we ran a K-nearest neighbour algorithm to cluster these tourist places into mainly 3 categories in order to get a sense of what kinds of destinations tend to stick together. The results obtained were quite astounding and gave us profound insights into what places are usually frequented together. The clusters were broadly significant of 3 main categories: **Women's/Fashion Stores, Restaurants & Coffee shops** and the third category going to the ones with finer tastes - **Deli, Arts & Crafts and Dance Studios.** These clusters could provide agencies with extremely valuable information about the visiting habits of tourists.

5. Discussions

The results obtained from this project gives us a lot to think about and work on when it comes to understanding the tourism industry of New York. The nuanced nature of tourist behaviour and what kinds of places they are most likely to visit is surely a topic of great interest for travel agencies in order to improve visitor experiences and company revenue alike.

The results showing how certain lesser known localities in New York are having some of the most frequented venues is definitely something to think about for tourist agencies when they are designing packages. This would certainly provide a very unique and authentic flavour of New York to any visitor who wishes to experience the essence of the city outside of the cliched descriptions that they see in advertisements. Such a package would surely add immense value that would eventually translate into increased revenue for any agency.

Another very interesting insight that the agencies can extract from this report is how different destinations tend to stick together and what it ultimately says about tourist behaviour and visiting habits. Different sets of visitors have diversified interests and given the right kind of data about their previous tourism behaviour, they can be matched to a package which suits their interests and presents them with a very happy surprise. Any experience like that would be worth millions to an agency which focuses on constantly improving visitor experiences to boost revenue.

6. Conclusion

This project was definitely a joy-ride which demonstrated how simple data about tourist's behaviour can help those involved in creating amazing experiences for their patrons. A lot of raw data still remains untapped in the tourism industry and efforts should be made to systematically collect, process and organize this data as it would surely reveal deeper insights into the tourism industry. The approach employed here can be extended for many major tourist destinations in the world like Paris, Rome, London, Mumbai, Tokyo etc. Thus, a lot still remains to be discovered and the future of the tourism industry looks data-bright!